# Industrial Internship Report On Project

# “Quality Prediction in a Mining Process”

# Domain: Data Science and Machine Learning

# Internship Duration: 15th June To 28th July

# Institutional Affiliation: UniConverge Technologies Pvt. Ltd. & UpSkill Campus

# Prepared By :- Date of Submission:-

# *Malkit Phogat* *28th July,2023*

# TABLE OF CONTENTS

|  |  |
| --- | --- |
| Preface | 3-5 |
| IntroductionAbout UniConverge Technologies Pvt Ltd.About upSkill CampusAbout IoT AcademyObjectiveReferences | 6-10 |
| Problem Statement | 11 |
| Existing and Proposed solution | 12-13 |
| Proposed Design/ ModelHigh Level DiagramLow Level DiagramInterfaces | 14-19 |
| Performance TestTest Plan/ Test CasesTest ProcedurePerformance Outcome | 20-29 |
| My Learning | 30-31 |
| Future Work Scope | 32-33 |

# 

# Preface

# In this 6 weeks industrial Internship organized by UniConverge Technologies Pvt. Ltd. In addition, UpSkill Campus and IoT Academy (Learning Resource Provider), I learnt a lot from my first industrial internship. UniConverge Company divide this internship in 6 weeks which give me a clear goal that how to deal with my problem statements which is given in my project.

# In First Week, we have to Explore our problem statements from our project which we choose from our own, actually 5-6 projects is given by UCT and we have to choose one of them and do exploring the problem statement. In addition, we have to learn about UniConverge Technologies (UCT) Company, that how it work, applications, technology used, etc.

# In Second Week, the project we choose we need to learn the instruction of the project and plan the best solution or model for the project problem statement. My domain is “Data Science and Machine Learning”. Therefore, I need to do lot of things on my dataset given by UCT like- Data Preprocessing, EDA, Modelling etc.

# In Third Week, I started working on my project and explore the dataset and find some tools from google, github, reference, etc., which give me clear path that which tools, is good for the project. I started the project with Python programming language, which is consider as the best programming language for Data Science and Machine Learning. I done some little bit of Data Preprocessing and Exploratory Data Analysis (EDA).

# In Fourth Week, I continued my working on my project and after Data Preprocessing and EDA; I started my other operations like- Outlier Removal, Resampling, Removal of Highly Correlated Features, Feature Engineering and Machine Learning Model. I done these things in my fourth week.

# In Fifth Week, After completing my project, in my fifth week I started make better performance of my model and like I tried to do add some other ML Algorithms which give me some great output from my previous model. In fifth week, I focus on model evaluation and improving the model performance.

# In Sixth Week, It is time for project submission and I my project is ready.

# Need of Relevant Internship in Career Development:

# Relevant internship are highly valuable for career development. This internship provide me hands-on experience, industry exposure, networking opportunities, skill development, resume enhancement, career exploration, and personal growth. Engaging in this internship can significantly contribute to my professional development and increase my chances of securing meaningful employment in Data Science and Machine Learning field.

# Problem Statement:

**• Is it possible to predict % Silica Concentrate every minute?**

In my first problem statement, I need to know that it is possible to predict %Silica Concentrate every minute, and when I applied the tools in my project it came to know that it is not possible because if we do it for every minute than lots of value is become null and data of the dataframe is also increased. It is difficult to find it in minutes that’s why we done it in hours.

**• How many steps (hours) ahead can we predict % Silica in Concentrate? This would help engineers to act in predictive and optimized way, mitigating the % of iron that could have gone to tailings.**

In this problem statement, we need to know that how much hours or time is needed to predict the %Silica Concentrate and when we done it our output is around 1817 hours. I done it and with this step our data become less and easy to handle and applying to create our model.

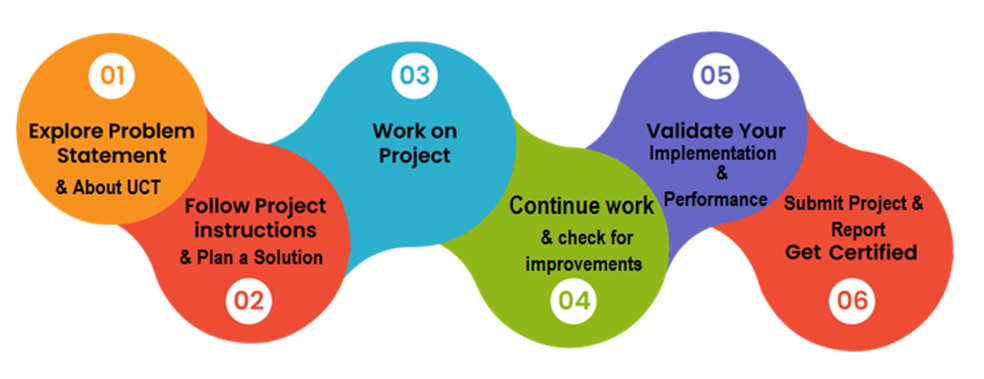
**• Is it possible to predict %Silica in Concentrate without using % Iron Concentrate column (as they are highly correlated)?**

In this problem statement, we need to know that is it possible to predict %Silica in Concentrate without using %Iron Concentrate column because this column is highly correlated and it will make easy to predict the accuracy. Therefore, we have to remove it and without this column, we have to make our model accurate.

**Opportunity given by UCT/USC:**

UCT/USC give me an opportunity to work on this amazing project during my internship, which give me a great exposure of industrial problems, and I learnt how to deal with all of these problems. I would like to thanks to UTC/USC who give me this opportunity that I will able to deal with real world problem and learn lot of things in this Data Science and Machine Learning field.

**How program was planned:**



**Learnings and overall Experience**:

This 6 weeks internship provide me a condensed yet valuable learning and overall experience. Some of the learning and experience I want to share which I learnt from this internship:

1. Practical Application of Knowledge
2. Skill Development
3. Industry Exposure
4. Networking Opportunity
5. Professional Growth and Self Awareness

Thank you to all staff of UpSkill Campus, UCT and IoT Academy who provide us such an amazing internship experience with their 24/7 hard work and support to complete this internship on time without any problem. I also want to thanks references, which help me to complete my project in proper way and on time. In our internship, there are three quizzes in alternate weeks and that is one of the great thing, which I experienced in my internship, quizzes help us to recall or do quick revision of the concept we learnt last week.

**Message to juniors and peers:**

First, embrace this opportunity with enthusiasm and an open mind. Your internship at UniConverge Technologies (UCT) will provide you with valuable insights into the industry and the chance to apply your academic knowledge in a practical setting. Take advantage of every task, project, and interaction to learn and grow both professionally and personally.

Networking is another crucial aspect of your internship. Connect with professionals in your field, attend company events, and engage in conversations. Building relationships can lead to mentorship, valuable advice, and potential future opportunities. Do not underestimate the power of a strong professional network.

I wish you all the best for your internship journey at UCT. Embrace every opportunity, be curious, and make the most of this valuable experience. If you have any questions or need support, feel free to reach out. I am confident that you will excel and have a fulfilling internship experience.

# Introduction

# About UniConverge Technologies Pvt Ltd.

# Founded in 2013, UniConverge Technologies has quickly established itself as a leading provider of innovative and high-quality digital solutions. Our team consists of talented and experienced professionals who are passionate about delivering exceptional results for our clients. With a focus on customer satisfaction and a commitment to excellence, we have built a reputation for being a trusted and reliable partner for businesses of all sizes. We have expertise in ‘Wireless Communication’ and ‘Internet of Things, product development and consulting services to companies working in Small Cells, Mobile Platforms, Healthcare, Medical Devices, Logistics, and Transportation and Manufacturing domains.

# Mission

# At UniConverge Technologies, our mission is to empower businesses to succeed in a rapidly evolving digital landscape. We do this by providing innovative and customized solutions that meet the unique needs of each of our clients. We believe in building long-lasting relationships with our clients based on trust, transparency, and mutual respect.

# 

# 

# 

# Vision

# Our vision is to offer organizations across the world, a wide gamut of services and solutions in the Wireless Communication and IOT domain. We are committed to continuously evolving and staying ahead of the curve to provide the best possible solutions to our clients.

# About upSkill Campus

# UpSkill Campus is a fast-growing ed-tech platform that is meant to upskill students, fresher’s, working professionals, faculties, entrepreneurs etc. We have the vision to provide an immersive experience for our learners to ensure their exhaustive growth. To provide learning 24x7 on the latest technologies that not only help to get a better job but also help in questioning and motivates you to do hands-on exercises and explore more.

# About IoT Academy

# The IoT Academy is an EdTech company imparting quality and industry-based programs for individuals, students, and working professionals. We offer exhaustive training in domains related to the Internet of Things (IoT), embedded systems, Data analytics, Industrial IoT, Big Data, Python, Artificial Intelligence, Machine Learning, and Industry 4.0. The IoT Academy is partnered with E&ICT Academy, IIT Kanpur, IIT Guwahati, IIT Roorkee for IoT skill development. Different programs are conducted to provide certification that is applicable in different countries like India, Africa, and the Middle East.

# Objective

The objective for this internship program was to

1. Get practical experience of working in the industry.
2. To solve real world problems.
3. To have improved job prospects.
4. To have improved understanding of our field and its applications.
5. To have Personal growth like better communication and problem solving.
   1. **References**
6. Google
7. GitHub
8. YouTube
9. Documents
10. Domain Professionals

# Problem Statements

# Introduction:

# The main goal is to use this data to predict how much impurity is present in the ore concentrate. The impurity is present in the form of Silica. Hence, the primary goal is to predict % Silica present in the Iron Ore concentrate. As this impurity is measured every hour, if we can predict how much silica (impurity) is in the ore concentrate, it can help the engineers, giving them early information to take actions and make process improvements. Hence, they will be able to take corrective actions in advance (reduce impurity, if it is the case) and help the environment (reducing the amount of ore that goes to tailings as you reduce silica in the ore concentrate).

# Problem Statements:

* Is it possible to predict % Silica Concentrate every minute?
* How many steps (hours) ahead can we predict % Silica in Concentrate? This would help engineers to act in predictive and optimized way, mitigating the % of iron that could have gone to tailings.
* Is it possible to predict % Silica in Concentrate without using % Iron Concentrate column (as they are highly correlated)?

# Dataset:

# The data is collected from machine sensors every 20 seconds and we have data from a 6-month range (March of 2017 until September of 2017). The dataset contains 737,453 instances and 24 attributes with the target variable being % Silica Concentrate. Our team decided to take two different approaches to see which approach will perform better. Due to the large dataset, we decided to create a resampled dataset that aggregated the sensor values to every 1 hour from 20 seconds, which became our second approach.

# Existing and Proposed Solution

# Proposed Solution Step by Step:

# Importing Useful Libraries for Data Analysis and Data Manipulation

# Numpy

# Pandas

# Matplotlib

# Seaborn

# Warnings\* (Optional)

# Data Preprocessing

# Data Analysis

# Exploratory Data Analysis

# Resampling of Data

# Identification of Outliers

# Outliers Removal

# Removal of Highly Correlated Features

# Feature Engineering

# Machine Learning Model

# Importing Scikit–Learn Machine Learning Libraries

# Train Test Split Data

# Creating Pipeline

# Cross Validation

# Train and Test Data

# Model Performance 1 (without tuning)

# Model Tuning Using GridSerachCV

# Final Model Performance 2 (with tuning)

# Value addition in my planning:

# As above roadmap of proposed solution, I will work same as I mention above. However, it is not much easy to analysis and perform machine-learning algorithms to predict the model performance or our problem statement. I need to review my project 2 to 3 times even after it completed. Lot of time my code getting error and need to spend lot of time to fix and continue my work again. Because, there are many machine-learning algorithms to deal with that type of data so, training and testing all of those is headache some time.

# Code Submission (GitHub Link):

# https://github.com/MalkitPhogat/MalkitPhogat.git

# Project Report Submission (GitHub Link):

# Proposed Model/Design and Solution

# Information of Given Dataset:

# Exploratory Data Analysis of Complete Dataset:

# Visualization for %Silica Concentrate w.r.t Time Period:

# EDA after Resampling Data and Removal of Outliers in Data:

# 

# EDA after applying Feature Engineering:

# 

# High Level Diagram:

# Therefore, High Level Diagram is my output or final model graphical representation where I will predict that my model is working approx. accurate. I split my model in two phases or in two ways:

# (First Phase): In this phase, I will take whole dataset, applied machine-learning algorithms, and check the model prediction. I did not apply any data removal, outliers removals, feature engineering etc. Just remove null values and convert data, which is generated every second. That phase is just for check the machine-learning algorithm in first phase. Although, this is not a final model. So, I want consider this diagram as my Low Level Diagram. (Not in Final Code which I submitted via GitHub link, it is just for experimental diagram.)

# (Second Phase): In this phase, I will take data after applying all methods like- resampling, outlier removal, feature engineering etc. After applying all these, size of the data is already reduced and then I take that data for my machine-learning model and make my final machine-learning model. Therefore, this diagram show my final model and I want to consider this diagram as my High Level Diagram.

# (Second Phase Diagram or High Level Diagram)

# (First Phase Diagram or Low Level Diagram)

# Test Performance

# The performance test conducted on the internship project aimed to predict the percentage of silica, which represents the impurity present in the ore concentrate. By accurately predicting the silica content, engineers can proactively take actions and make process improvements, leading to several benefits.

# Primarily, predicting the silica content every hour provides engineers with early information about the impurity levels. This allows them to identify any deviations or anomalies promptly and take immediate corrective actions. By addressing impurity issues early on, the engineers can prevent further processing of ore concentrate with high silica content, thereby reducing the production of substandard or unusable materials.

# The performance test results provide crucial insights into the accuracy and reliability of the silica prediction model. Engineers can assess the model's performance in terms of response time, throughput, scalability, and resource utilization. This information helps in determining the efficiency of the prediction process and identifies any bottlenecks or areas for improvement.

# By leveraging the data obtained from the performance test, engineers can confidently rely on the silica prediction model to make informed decisions and take proactive measures to optimize the ore concentration process. Ultimately, this contributes to higher production efficiency, improved product quality, and reduced environmental impact, aligning with the project's objectives of process optimization and sustainability.

# Test Plan/Test Cases

# After applying Data Analysis, Data Manipulation, Exploratory Data Analysis (EDA), Feature Engineering part it is time to make Machine Learning Model.

# Therefore, now we have to train and test our data. After splitting our data, it is time to test it, check the accuracy, and predict the target variable.

# Now, before testing the data in any randomly picked machine-learning algorithm we have to check which machine-learning algorithm is best fitted for our project. However, our project problem is “Regression Problem” which is a supervised learning based problem. Therefore, we have to apply those machine-learning algorithms, which is work good for our project problem.

# Plan for Machine Learning Model:

# Some Machine Learning Algorithms we test to check the performance of each of one algorithm are:

# Linear Regression (LR)

# Stochastic Gradient Descent Regression (SGDR)

# Decision Tree Regression (DTR)

# Support Vector Regression (SVR)

# Random Forest Regression (RFR)

# Gradient Boosting Regression (GBR)

# K Nearest Neighbors (KNN)

# I will use all these supervised machine-learning algorithms to test the performance of all the algorithms for our data.

# Creating Pipelines:

# The purpose of creating a pipeline in a machine learning (ML) model is to streamline and automate the data processing and model training workflows. A pipeline enables a systematic and efficient approach to building, deploying, and maintaining ML models.

# Calculate Root Mean Squared Error (RMSE) of all used machine-learning algorithms (This process is known as cross validation):

# It is important for an algorithm to give us less error as possible. Because, errors will effect on our machine learning model and reduce the model accuracy that is going to predict wrong prediction in future.

# Test Data for Validate Model:

# The purpose of testing data to validate a model in machine learning is to assess the performance and generalization capability of the model on unseen data. It involves evaluating how well the model can make accurate predictions or classifications on data that it has not been trained on.

# Model Tuning Using GridSearchCV:

# The purpose of model tuning using GridSearchCV in machine learning is to systematically search for the best combination of hyperparameters for a given model. GridSearchCV is a technique that exhaustively searches through a specified grid of hyperparameter values, evaluating the model's performance for each combination.

# Final Validation of Model based on Best Estimator, which we obtain using through GridSearchCV:

# The purpose of the final validation of a model based on the best estimator obtained through GridSearchCV is to assess the model's performance on a completely independent dataset. This final validation step provides a reliable estimate of how well the model is expected to perform in real-world scenarios

# Test Procedure

# Importing relevant Scikit Learn Libraries:

# train\_test\_split

# Pipline

# StandardScaler

# LinearRegressor

# SGDRegressor

# DecisionTreeRegressor

# SVR

# RandomForestRegressor

* GradientBoostingRegressor

# KNeigbhorsRegressor

# Metrics

# Splitting Data into Training and Testing

# Creating Piplines

# Cross Validation and KFold (With Scikit Learn Packages)

# cross\_val\_score

# KFold

# So, to overcome with errors we will calculate RMSE (Root Mean Squared Error) which give us the information that how much that particular model produce the difference between predicted and actual value. Therefore, those models who has minimum RMSE that will going to predict most accurate prediction.

# Transform the data with “StandardScaler()”, fit the training data into the machine-learning model and calculate “RMSE” and “R2” score.

# Predict the model performance (without model tuning).

# Tuning our model using GridSearchCV to find the best parameter for our model that predict more accurate output from previous one.

# Again, train and test the model with those best parameters that we obtain from above step and calculate the “RMSE” and “R2” score.

# Again, predict the model performance (with model tuning).

# Check the improvement/difference between model (without tuning) and model (with tuning). Check how that our model improve in its accuracy from previous model accuracy which we done without model tuning.

# Performance Outcome

# Our outcome is based on our problem statements. Therefore, I will show my outcomes in three part, because we have three-problem statement and our main goal is to find the solution for those problem statements.

# Is it possible to predict % Silica Concentrate every minute?

# As you are able to saw above data, when we want to predict %Silica Concentrate every minute then most of the data is showing null. The reason behind that is, when we calculate for every minute then lots of value is become NaN or empty it means they are missing value that it is given that the sensors are record the data for every 20 seconds and it is difficult to get data for every minute.

# Therefore, it is not possible to predict the %Silica Concentrate every minute.

# Solution of the Problem:

# To overcome with problem we will calculate the data for every hour instead of every minute, which make a sense because if we do it then it became easy to calculate a mean of all value that come in one hour.

# As you notice from above table, when we calculate or try to predict for every hour we will able to obtain the value without any null or missing values.

# Therefore, we will able to predict %Silica Concentrate every hour instead of every minute.

# How many steps (hours) ahead can we predict % Silica in Concentrate? This would help engineers to act in predictive and optimized way, mitigating the % of iron that could have gone to tailings.

# After find the solution of previous problem statement our next problem statement said that, how much time (hours) ahead/need to predict %Silica Concentrate.

# Therefore, after converting our dataframe time from every 20 seconds to every hours it’s time to analyze that how much time it will take. But, we cannot do this step just after applying previous step, because still there are lots of data is missing in our dataframe (as given in code) and there are lots of outliers present in dataframe and we need to remove all those to obtain is clean data to make machine learning model. Therefore, we need to remove all those rows.

# 

# After removing all missing values and outliers, we obtain our new data, which consist 1817 rows and 23 columns.

Hence, there are 1817 rows in dataframe. Therefore, that means we are going to predict % Silica Concentrate in 1817 hours, which is around 75 days (2.5 months).

Therefore, That’s mean we are going to predict the % Silica Concentrate for 1817 hours after resampling, filtering and removing null or unused data’s.

# Is it possible to predict % Silica in Concentrate without using % Iron Concentrate column (as they are highly correlated)?

# In this problem statement, we need to first remove %Iron Concentrate column (which is highly correlated) and then apply machine learning algorithms to predict the %Silica in Concentrate.

# After removing the %Iron Concentrate column we need to do some feature engineering (as given in code) and after that we are able apply machine-learning algorithms.

# Machine Learning Model

# Selection of machine-learning algorithm based on RMSE value:

# Therefore, from above graph we are able conclude that “Random Forest Regressor” (RMSE = 0.941 which is minimum from others) is the best machine-learning algorithm for our dataset to predict the %Silica Concentrate.

# Calculate the RMSE and R2 of predicted values, which we obtain from our selected machine-learning algorithm (RFR):

# RMSE: 0.911

# R2: 0.224

# Performance of the Model:

# Accuracy Error: 0.707

# Accuracy of the Model: 62.358%

# 

# To improve our model performance we need to apply GridSearchCV to obtain best parameters for our model and then we again check performance of our model:

# Model Performance after tuning our model with best parameters:

# RMSE: 0.900

# R2: 0.243

# Accuracy Error: 0.696

# Accuracy of the Model: 62.786%

# Improvement: 0.43%

# Scatter Plot between Y Test (Actual Data) vs Y Prediction (Predicted Data)

# According to graph or plot, we finally conclude that, Yes, it is possible to predict the %Silica in Concentrate without %Iron Concentrate column.

# Conclusion:

# After using various algorithms, I documented their performance using the RMSE score, R2, and accuracy. I found out that the Random Forest Regressor performed better on the resampled dataset that aggregated the values every hour from the original dataset. I were able to predict the % Silica Concentrate with an accuracy 62.786% on the small resampled dataset.

# My Learning

# During my 6-week internship in the field of data science and machine learning, I gained valuable knowledge and experiences. Here are some key takeaways from my internship:

1. Data preprocessing: I learned the importance of data cleaning and preprocessing. This involved handling missing values, outliers, and transforming variables to ensure the data was suitable for analysis.
2. Exploratory data analysis (EDA): I gained hands-on experience in performing EDA techniques to gain insights into the data. This included data visualization, statistical analysis, and identifying patterns and trends in the data.
3. Machine learning algorithms: I had the opportunity to work with various machine-learning algorithms such as linear regression, logistic regression, decision trees, random forests, and support vector machines. I learned how to train, validate, and evaluate these models for predictive and classification tasks.
4. Feature engineering: I learned about the importance of feature engineering in improving model performance. This involved creating new features, selecting relevant features, and transforming variables to enhance the predictive power of the models.
5. Model evaluation and selection: I gained insights into different evaluation metrics for regression and classification tasks. I learned how to assess model performance using metrics such as accuracy, precision, recall, F1 score, and mean squared error.
6. Hyperparameter tuning: I got hands-on experience in optimizing model performance by tuning hyperparameters. This involved techniques such as grid search, random search, and using cross-validation to find the best combination of hyperparameters.
7. Model deployment: I learned about deploying machine-learning models in real-world applications. This involved considerations such as model serialization, model serving, and integrating the model into production systems.
8. Connect with Professional: Important aspect of my internship experience in data science and machine learning was the opportunity to establish professional connections. Throughout the internship, I had the chance to interact and collaborate with professionals in the field, including mentors, supervisors, and colleagues.

Overall, my internship in data science and machine learning provided me with a solid foundation in understanding and applying key concepts and techniques in this field. It enhanced my practical skills and gave me valuable insights into real-world data analysis and modeling scenarios.

**8. Future Work Scope**

Based on the internship experience I mentioned above in the field of data science and machine learning, there are several potential future work scopes I can consider:

1. Further Research: I can use the knowledge and skills acquired during the internship to delve deeper into specific topics or areas of interest within data science and machine learning. This could involve exploring advanced machine learning algorithms, investigating cutting-edge research papers, or studying specific domains where data science techniques can be applied.
2. Advanced Projects: Undertake more complex and challenging projects to expand my practical experience. This could involve working on larger datasets, tackling real-world problems, or implementing advanced techniques such as deep learning or natural language processing.
3. Specialization: Consider specializing in a specific subfield within data science and machine learning. This could involve focusing on areas such as computer vision, natural language processing, or time series analysis. Developing expertise in a specific area can enhance my career prospects and open up niche opportunities.
4. Industry Applications: Apply my knowledge and skills to real-world applications in different industries. Data science and machine learning techniques have a wide range of applications across sectors like finance, healthcare, e-commerce, and marketing. Explore opportunities to work on projects or internships in specific industries to gain domain-specific knowledge and experience.
5. Continued Learning: Data science and machine learning are rapidly evolving fields, with new techniques and technologies emerging regularly. Stay updated with the latest developments by regularly reading research papers, participating in online courses or webinars, and joining data science communities. Continuous learning will help me stay competitive and adapt to new challenges.
6. Networking and Collaboration: Continue building professional connections in the field. Attend conferences, join industry forums or meetups, and engage in online communities to network with professionals and researchers. Collaborate on projects, share knowledge, and stay connected with the latest trends and opportunities.
7. Advanced Degree or Certification: Consider pursuing an advanced degree or certification in data science or a related field. This can provide me with a more comprehensive understanding of the subject, access to advanced coursework, and potentially open doors to higher-level positions or research opportunities.

Top of Form

Top of Form

# 